

Class 19: Pertussis Mini Project

Junlin Ruan (PID: A17839687)

Background

Pertussis is a bacterial lung infection also known as Whooping cough. Let's begin by examining CDC reported case numbers in the US:

```
cdc <- data.frame(
  year = c(1922L, 1923L, 1924L, 1925L,
           1926L, 1927L, 1928L, 1929L, 1930L, 1931L,
           1932L, 1933L, 1934L, 1935L, 1936L,
           1937L, 1938L, 1939L, 1940L, 1941L, 1942L,
           1943L, 1944L, 1945L, 1946L, 1947L,
           1948L, 1949L, 1950L, 1951L, 1952L,
           1953L, 1954L, 1955L, 1956L, 1957L, 1958L,
           1959L, 1960L, 1961L, 1962L, 1963L,
           1964L, 1965L, 1966L, 1967L, 1968L, 1969L,
           1970L, 1971L, 1972L, 1973L, 1974L,
           1975L, 1976L, 1977L, 1978L, 1979L, 1980L,
           1981L, 1982L, 1983L, 1984L, 1985L,
           1986L, 1987L, 1988L, 1989L, 1990L,
           1991L, 1992L, 1993L, 1994L, 1995L, 1996L,
           1997L, 1998L, 1999L, 2000L, 2001L,
           2002L, 2003L, 2004L, 2005L, 2006L, 2007L,
           2008L, 2009L, 2010L, 2011L, 2012L,
           2013L, 2014L, 2015L, 2016L, 2017L, 2018L,
           2019L, 2020L, 2021L, 2022L, 2023L, 2024L),
  cases = c(107473, 164191, 165418, 152003,
            202210, 181411, 161799, 197371,
            166914, 172559, 215343, 179135, 265269,
            180518, 147237, 214652, 227319, 103188,
            183866, 222202, 191383, 191890, 109873,
            133792, 109860, 156517, 74715, 69479,
            120718, 68687, 45030, 37129, 60886,
```

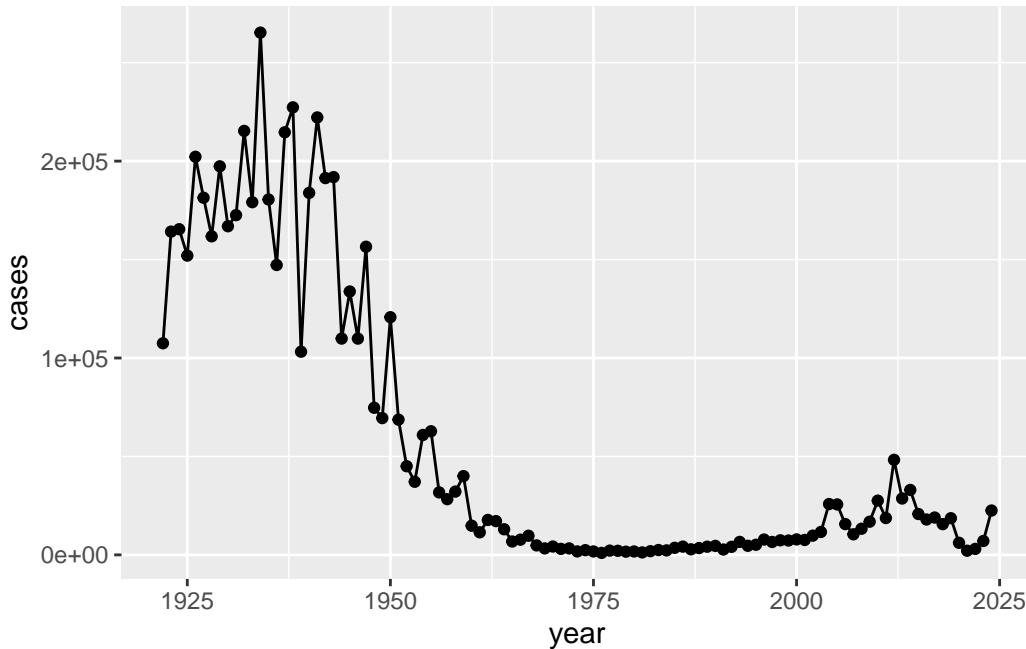
```
)
62786,31732,28295,32148,40005,
14809,11468,17749,17135,13005,6799,
7717,9718,4810,3285,4249,3036,
3287,1759,2402,1738,1010,2177,2063,
1623,1730,1248,1895,2463,2276,
3589,4195,2823,3450,4157,4570,
2719,4083,6586,4617,5137,7796,6564,
7405,7298,7867,7580,9771,11647,
25827,25616,15632,10454,13278,
16858,27550,18719,48277,28639,32971,
20762,17972,18975,15609,18617,
6124,2116,3044,7063, 22538)
```

Plot of cases per year for Pertussis in the US

Q1. With the help of the R “addin” package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
library(ggplot2)

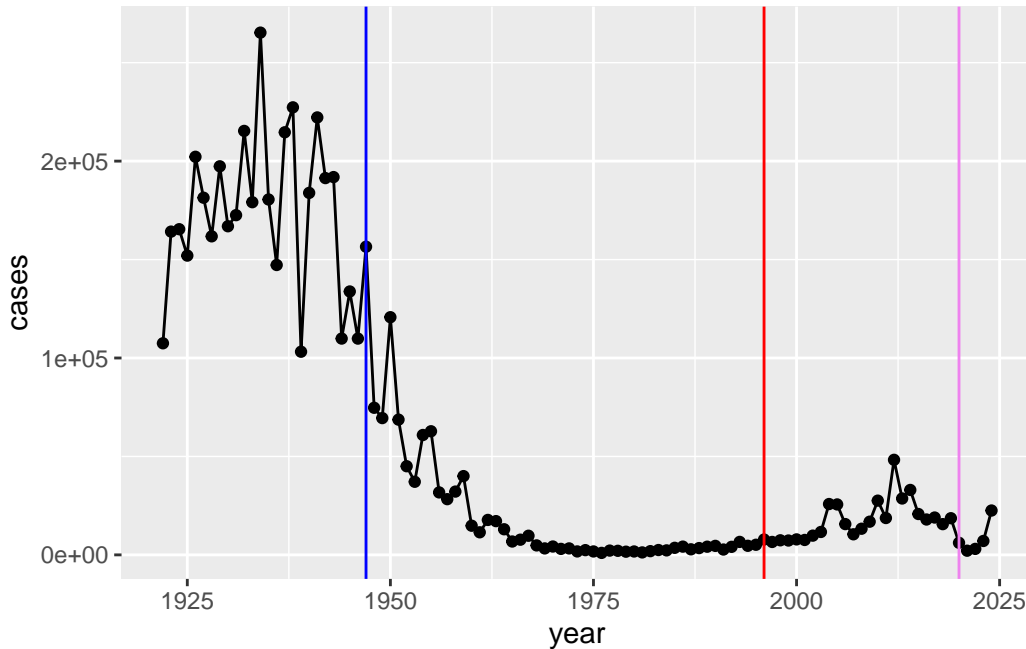
ggplot(cdc) +
  aes(year, cases) +
  geom_point() +
  geom_line()
```



Q2. Using the ggplot `geom_vline()` function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

Add some major milestone time points to our plot:

```
ggplot(cdc) +
  aes(year, cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1947, col = "blue") +
  geom_vline(xintercept = 1996, col = "red") +
  geom_vline(xintercept = 2020, col = "violet")
```



The full introduction of the wP (whole-cell) Pertussis immunization in the mid 1940s lead to dramatic reduction in case numbers (from over 200,000 to 100s).

Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

The switch to the aP (newer acellular formalization in the US did not result in further decline to the case number but increase of case number shown after the switch.

The 2020 lock-down and social distancing measures help mitigate the propagation of the disease, resulted in temporary decline of case number.

The CMI-PB Project

The mission of CMI-PB is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of Pertussis booster vaccination.

Website: <https://www.cmi-pb.org/>

They make their data available via JSON format API endpoints - basically the database tables in a key:value type format like "infancy_vac": "wP". To read this we can use the `read_json()` function from the `jsonlite` package by install with `install.packages("jsonlite")`.

```
library(jsonlite)

subject <- read_json(path = "https://www.cmi-pb.org/api/v5_1/subject",
                      simplifyVector = TRUE)

head(subject)
```

	subject_id	infancy_vac	biological_sex	ethnicity	race
1	1	wP	Female	Not Hispanic or Latino	White
2	2	wP	Female	Not Hispanic or Latino	White
3	3	wP	Female	Unknown	White
4	4	wP	Male	Not Hispanic or Latino	Asian
5	5	wP	Male	Not Hispanic or Latino	Asian
6	6	wP	Female	Not Hispanic or Latino	White

	year_of_birth	date_of_boost	dataset
1	1986-01-01	2016-09-12	2020_dataset
2	1968-01-01	2019-01-28	2020_dataset
3	1983-01-01	2016-10-10	2020_dataset
4	1988-01-01	2016-08-29	2020_dataset
5	1991-01-01	2016-08-29	2020_dataset
6	1988-01-01	2016-10-10	2020_dataset

Q. How many “subjects”/individuals are in this dataset?

```
nrow(subject)
```

```
[1] 172
```

Q4. How many wP and aP subjects are there?

```
table(subject$infancy_vac)
```

```
aP wP
87 85
```

Q5/6. What is the breakdown by “biological_sex” and “race”?

```
table(subject$biological_sex)
```

Female	Male
112	60

```
table(subject$race)
```

American Indian/Alaska Native	1
Asian	44
Black or African American	5
More Than One Race	19
Native Hawaiian or Other Pacific Islander	2
Unknown or Not Reported	21
White	80

```
table(subject$race, subject$biological_sex)
```

	Female	Male
American Indian/Alaska Native	0	1
Asian	32	12
Black or African American	2	3
More Than One Race	15	4
Native Hawaiian or Other Pacific Islander	1	1
Unknown or Not Reported	14	7
White	48	32

This breakdown is not particularly representative of the US population - this is a serious caveat for this study. However, it is still the largest sample of it's type every assembled.

```
specimen <- read_json("https://www.cmi-pb.org/api/v5_1/specimen",
                      simplifyVector = TRUE)
ab_titer <- read_json("https://www.cmi-pb.org/api/v5_1/plasma_ab_titer",
```

```
simplifyVector = TRUE)

head(specimen)
```

	specimen_id	subject_id	actual_day_relative_to_boost	
1	1	1		-3
2	2	1		1
3	3	1		3
4	4	1		7
5	5	1		11
6	6	1		32

	planned_day_relative_to_boost	specimen_type	visit
1	0	Blood	1
2	1	Blood	2
3	3	Blood	3
4	7	Blood	4
5	14	Blood	5
6	30	Blood	6

Q9. Complete the code to join specimen and subject tables to make a new merged data frame containing all specimen records along with their associated subject details:

We need to “join” or link these tables with the `subject` table so we can begin to analyze this data and know who a given Ab sample was collected for and when.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.4      v readr      2.1.5
v forcats    1.0.1      v stringr    1.5.2
v lubridate  1.9.4      v tibble     3.3.0
v purrr      1.1.0      v tidyr      1.3.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x purrr::flatten() masks jsonlite::flatten()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
meta <- inner_join(subject, specimen)
```

Joining with `by = join_by(subject_id)`

```
head(meta)
```

	subject_id	infancy_vac	biological_sex	ethnicity	race
1	1	wP	Female	Not Hispanic or Latino	White
2	1	wP	Female	Not Hispanic or Latino	White
3	1	wP	Female	Not Hispanic or Latino	White
4	1	wP	Female	Not Hispanic or Latino	White
5	1	wP	Female	Not Hispanic or Latino	White
6	1	wP	Female	Not Hispanic or Latino	White

	year_of_birth	date_of_boost	dataset	specimen_id
1	1986-01-01	2016-09-12	2020_dataset	1
2	1986-01-01	2016-09-12	2020_dataset	2
3	1986-01-01	2016-09-12	2020_dataset	3
4	1986-01-01	2016-09-12	2020_dataset	4
5	1986-01-01	2016-09-12	2020_dataset	5
6	1986-01-01	2016-09-12	2020_dataset	6

	actual_day_relative_to_boost	planned_day_relative_to_boost	specimen_type
1	-3	0	Blood
2	1	1	Blood
3	3	3	Blood
4	7	7	Blood
5	11	14	Blood
6	32	30	Blood

	visit
1	1
2	2
3	3
4	4
5	5
6	6

Now let's join the `ab_titer` table with our `meta` table so we have all information about a given Ab measurement

```
ab_data <- inner_join(meta, ab_titer)
```

Joining with `by = join_by(specimen_id)`

```
head(ab_data)
```

	subject_id	infancy_vac	biological_sex	ethnicity	race
1	1	wP	Female Not Hispanic or Latino	White	
2	1	wP	Female Not Hispanic or Latino	White	
3	1	wP	Female Not Hispanic or Latino	White	
4	1	wP	Female Not Hispanic or Latino	White	
5	1	wP	Female Not Hispanic or Latino	White	
6	1	wP	Female Not Hispanic or Latino	White	

	year_of_birth	date_of_boost	dataset	specimen_id
1	1986-01-01	2016-09-12	2020_dataset	1
2	1986-01-01	2016-09-12	2020_dataset	1
3	1986-01-01	2016-09-12	2020_dataset	1
4	1986-01-01	2016-09-12	2020_dataset	1
5	1986-01-01	2016-09-12	2020_dataset	1
6	1986-01-01	2016-09-12	2020_dataset	1

	actual_day_relative_to_boost	planned_day_relative_to_boost	specimen_type
1	-3	0	Blood
2	-3	0	Blood
3	-3	0	Blood
4	-3	0	Blood
5	-3	0	Blood
6	-3	0	Blood

	visit	isotype	is_antigen_specific	antigen	MFI	MFI_normalised	unit
1	1	IgE	FALSE	Total	1110.21154	2.493425	UG/ML
2	1	IgE	FALSE	Total	2708.91616	2.493425	IU/ML
3	1	IgG	TRUE	PT	68.56614	3.736992	IU/ML
4	1	IgG	TRUE	PRN	332.12718	2.602350	IU/ML
5	1	IgG	TRUE	FHA	1887.12263	34.050956	IU/ML
6	1	IgE	TRUE	ACT	0.10000	1.000000	IU/ML

	lower_limit_of_detection
1	2.096133
2	29.170000
3	0.530000
4	6.205949
5	4.679535
6	2.816431

Q. How many Ab measurements do we have in total

```
nrow(ab_data)
```

```
[1] 61956
```

Q. How many different isotypes (types of antibody(Ab))?

```
unique(ab_data$isotype)
```

```
[1] "IgE" "IgG" "IgG1" "IgG2" "IgG3" "IgG4"
```

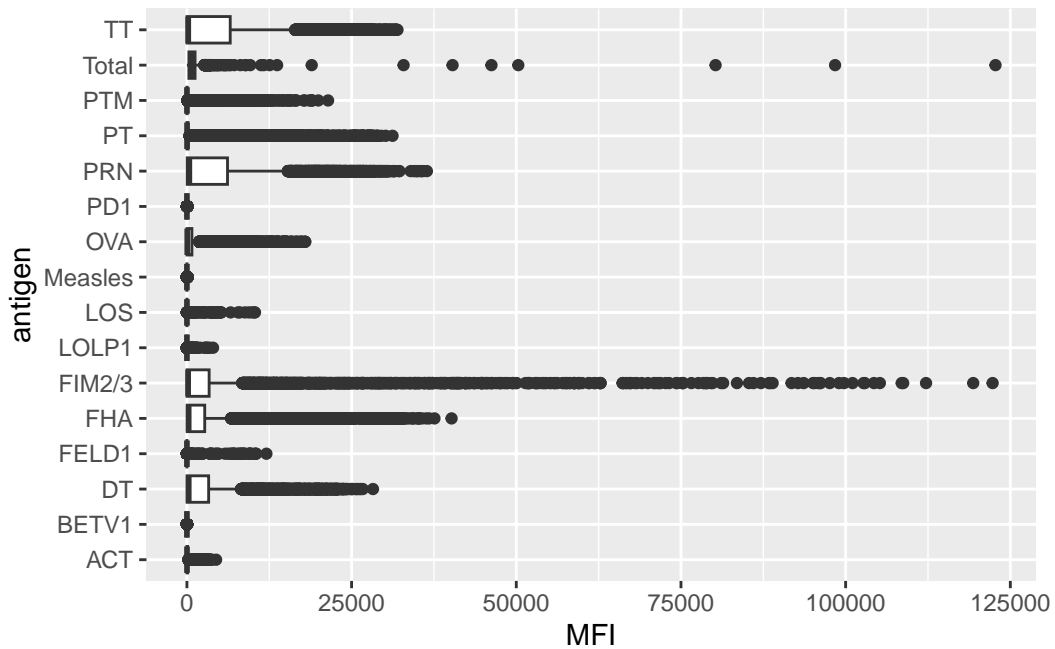
Q. How many different antigens?

```
unique(ab_data$antigen)
```

```
[1] "Total" "PT" "PRN" "FHA" "ACT" "LOS" "FELD1"  
[8] "BETV1" "LOLP1" "Measles" "PTM" "FIM2/3" "TT" "DT"  
[15] "OVA" "PD1"
```

```
ggplot(ab_data) +  
  aes(MFI, antigen) +  
  geom_boxplot()
```

Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).

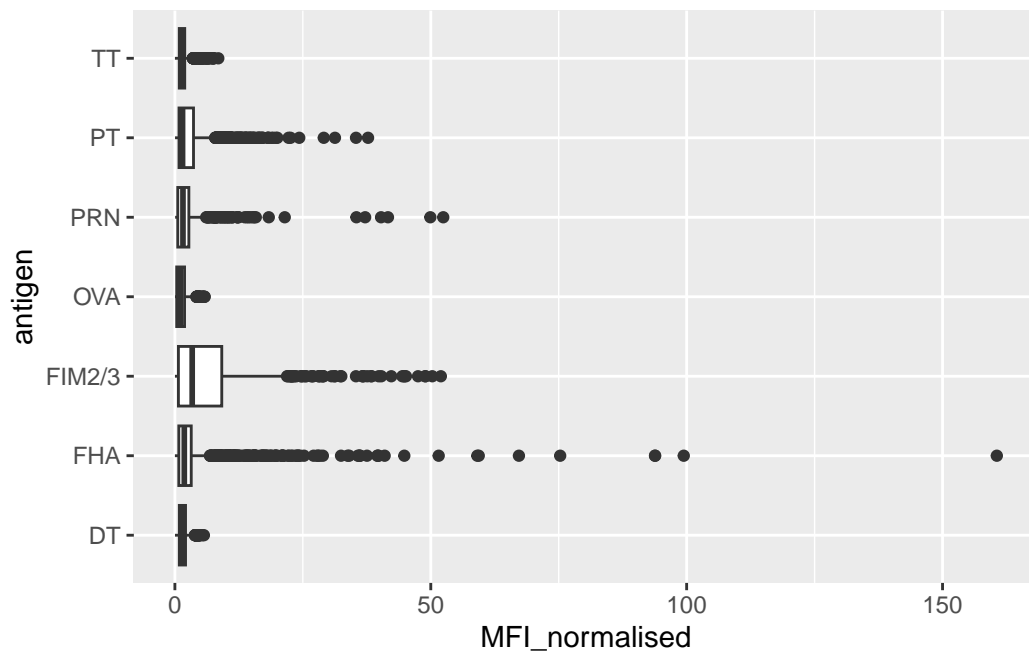


Examine IgG Ab titer levels

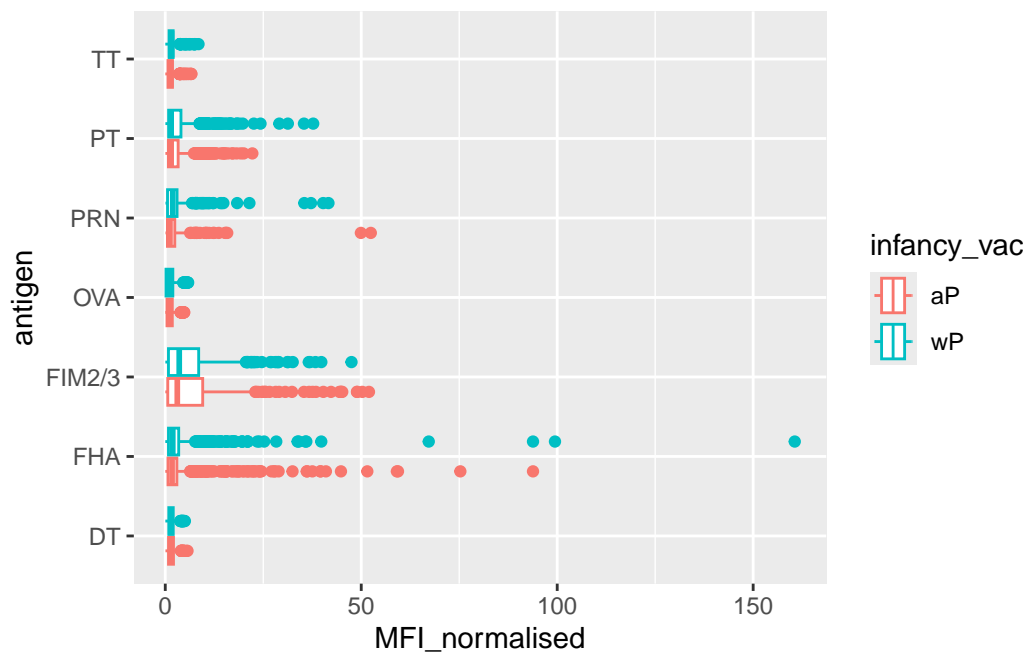
IgG is crucial for long-term immunity and responding to bacterial & viral infections

```
igg <- ab_data |>
  filter(isotype == "IgG")
```

```
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot()
```

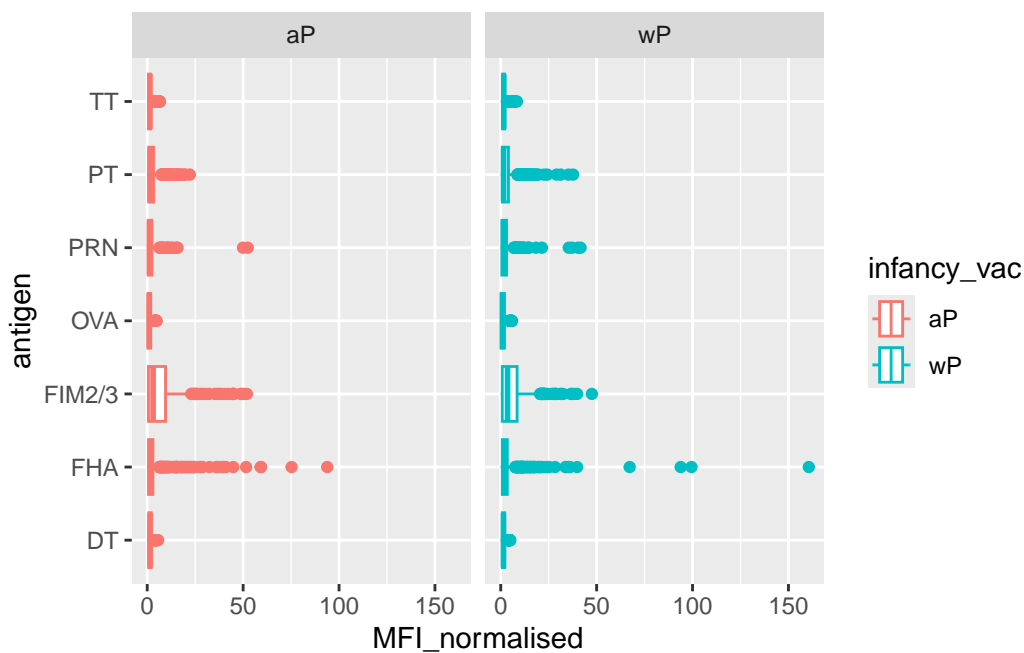


```
ggplot(igg) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot()
```



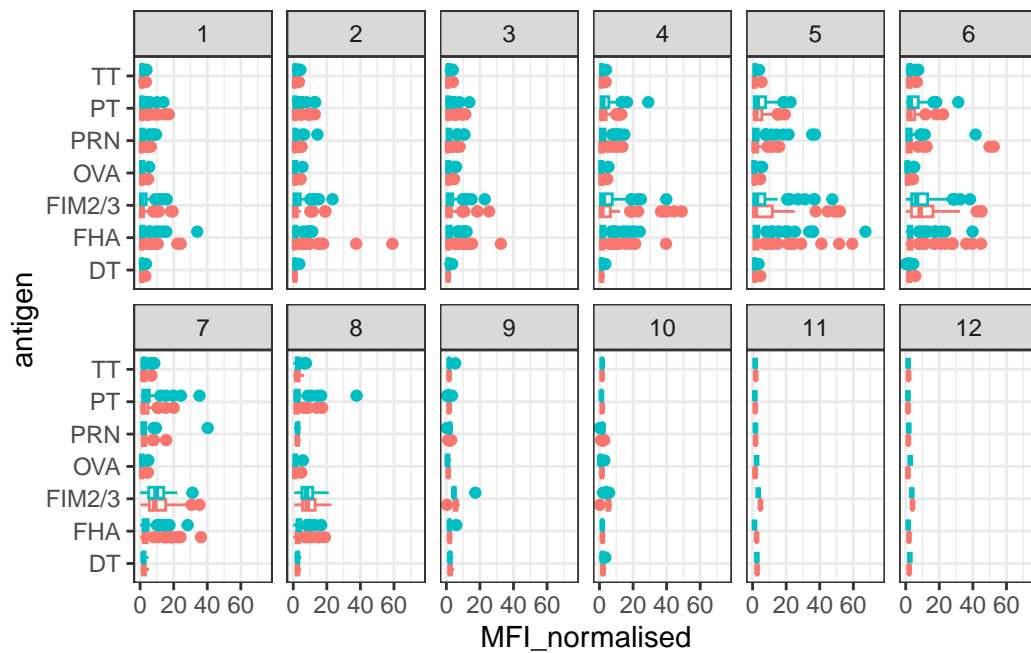
We can “facet” our plot by wP vs aP

```
ggplot(igg) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(~infancy_vac)
```



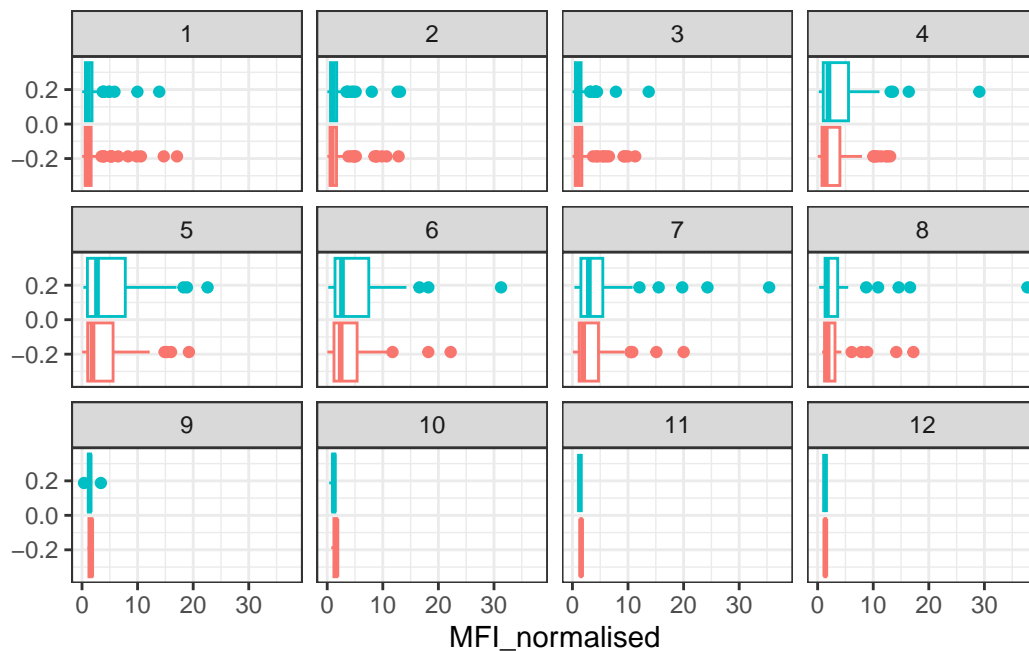
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

Warning: Removed 5 rows containing non-finite outside the scale range (`stat_boxplot()`).



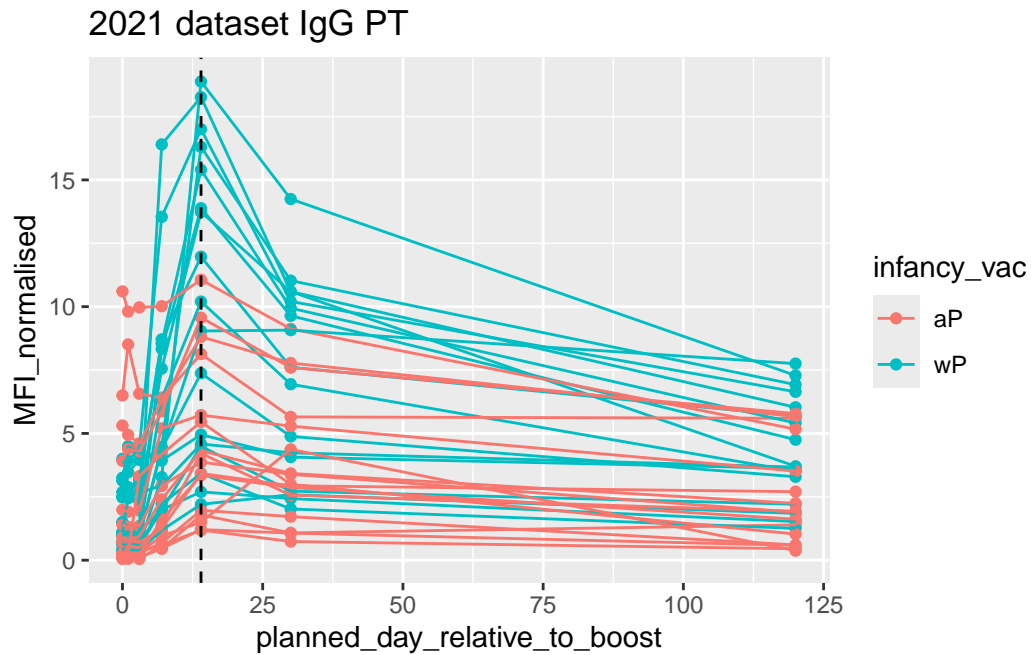
More advanced analysis digging into individual antigen responses over time:

```
filter(igg, antigen=="PT") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Lets finish this section by looking at the 2021 dataset IgG PT antigen levels time-course:

```
filter(igg, antigen == "PT", dataset == "2021_dataset") %>%
  ggplot() +
    aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,
        group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=14, linetype="dashed") +
    labs(title="2021 dataset IgG PT")
```



This plot shows the time course of Pertussis toxin (PT) antibody responses for a large set of wP (teal color) and aP (red color) individuals. Levels peak at day 14 and are larger in magnitude for wP than aP individuals.

There are lots of cool things to explore in this dataset and we need coding and biology knowledge to do it effectively - i.e. us!